Mobile Data Traffic Modeling: Revealing Temporal Facets

Eduardo Mucelli Rezende Oliveira, Aline Carneiro Viana, Kolar Purushothama Naveen, Carlos Sarraute

Mobile Data Traffic Modeling: Revealing Temporal Facets

Eduardo Mucelli Rezende Oliveira[a,b,*], Aline Carneiro Viana[b], K. P. Naveen[b], Carlos Sarraute[c]

[a] *École Polytechnique, France*
[b] *INRIA, France*
[c] *Grandata Labs, Argentina*

Abstract

This paper presents a detailed measurement-driven model of mobile data traffic usage of smartphone subscribers, using a large-scale dataset collected from a major 3G network in a dense metropolitan area. Our main contribution is a synthetic, measurement-based, mobile data traffic generator capable of simulating traffic-related activity patterns over time for different categories of subscribers and time periods for a typical day in their lives. We first characterize individual subscribers' routinary behaviour, followed by a detailed investigation of subscribers' temporal usage patterns (i.e., "when" and "how much" traffic is generated). We then classify the subscribers into six distinct profiles according to their usage patterns and model these profiles according to two daily time periods: peak and non-peak hours. We show that the synthetic trace generated by our data traffic model consistently replicates a subscriber's profiles for these two time periods when compared to the original dataset. Broadly, our observations bring important insights into temporal network resource usage. We also discuss relevant issues in traffic demands and describe implications in network solution evaluation and privacy.

Keywords: traffic analysis, mobile data, synthetic traffic, clustering

1. Introduction

Smartphone devices provide today the best means of gathering users' information about content consumption behavior on a large scale. In this context, the literature presents efforts in order to study and model the mobility of users, but little is publicly known about users content consumption patterns. The *understanding of users' mobile data traffic demands* is of fundamental importance when looking for solutions to manage the recent boost of mobile data usage [1, 2, 3] and to improve the quality of communication services provided, favoring the proliferation of pervasive communication. Hence, the definition of a *temporal usage pattern* can allow telecommunication operators to better foresee traffic demands, and consequently, to better timely plan network resources allocation and subsequently set subscription plans.

Contrarily to a significant number of related work using voice call and text message information from Call Detail Records (CDRs), we characterize and model data traffic demand generated by smartphone subscribers. Although convenient and frequently used, voice calls and SMS

*Corresponding author
Email address: edumucelli@inria.fr (Eduardo Mucelli Rezende Oliveira)

records only provide an approximation of users' data consumption. In addition, due to its sparsity in time [4], subscribers calling behavior shows strong variations respect to the time and day of the week [3], which is not the case for data traffic. Also, call traffic misses the background traffic load automatically generated by current smartphone applications (e.g., email checks, synchronization). Most importantly, since smartphones are now used more for data than for calls [5], the use of call records for investigating traffic demands is not enough for dimensioning network usages.

The contributions of this paper are threefold. It provides *a detailed analysis of the behavior of mobile subscribers in terms of the traffic they generate, from both a network and social perspective, using a large-scale dataset from a major telecom operator in Mexico City*. The outcomes of the analysis in terms of traffic and temporal dynamics gave us the required insights to better set modeling parameters, such as the period and the type of traffic (i.e., upload vs download) to be considered. In particular, thanks to the analysis, we could identify the high correlation between the upload and download volumes. This induced us to take into consideration the total volume per session (i.e., the sum of the upload and download volumes during the session) in our evaluation and traffic modeling. We have also detected the similarity of the temporal activity patterns among different days of the week, what brought us to the decision of modeling only 1 day of traffic, what we verified to be enough to profile subscribers traffic behavior. Second, we *profile urban mobile data traffic over time*. For this, we perform a *precise temporal characterization of individual subscribers' traffic behavior clustered by their usage patterns*, instead of a network-wide data traffic view [6, 7, 8]. Note that the high variability of individual subscribers' traffic demands and the use of large scale datasets make this task complex. Our third contribution is to *provide a way to synthetically, still consistently, reproduce temporal usage patterns of mobile subscribers* (the first work in the literature to do so, to the best of our knowledge). The implications of this work are diverse, in particular, in temporal analysis of network resource allocation. When it comes to privacy issues, it is worth mentioning that synthetic datasets do not risk exposing subscribers' private information, and may be used by any entity willing to perform realistic network simulations.

We perform our study on an anonymized dataset collected at the core of a major 3G network of Mexico's capital (Section 3). The dataset spans 4 consecutive months from July to October 2013 and consists of all the data traffic generated by 6.8 million subscribers. The dataset contains detailed information on the volume and frequency of all the data traffic generated by smartphone subscribers. This includes any uploaded and downloaded data, i.e., not only browsing traffic, but also the traffic automatically generated by applications. This represents an order of thousands of terabytes exchanged in the largest city of Mexico. Moreover, the dataset provides information about age and gender for more than half a million subscribers allowing us to segment our analysis along these demographic attributes.

We first analyse traffic usage habits as a function of time, age, and gender (Section 3). We observe similar usage patterns on different days. This allows us to choose a single day for studying the subscribers' temporal usage pattern (i.e., "when" and "how much" they generate traffic) in detail. In order to consistently analyse the heterogeneous usage of a large number of subscribers, we classify them into six distinct profiles according to their usage pattern (Section 4). We finally model the usage pattern of these six subscriber profiles according to two different journey periods: peak and non-peak hours. Using a sample dataset and numerous statistical tools, we show that our traffic model is able to consistently imitate different subscribers profiles over time in two journey periods, when compared to the original traffic dataset (Section 5). Our main result is *a synthetic, measurement-based mobile data traffic generator, capable of imitating traffic-related activity patterns in a temporal scale for six different categories of subscribers, during two time*

periods of a routinary normal day in their lives. We discuss the implications of our contributions in Section 6 and related work in Section 2. Finally, Section 7 concludes this paper. In this paper, user and subscriber will be used interchangeably.

2. Related Work

The understanding of users' content consumption has attracted significant attention of the networking community in the literature. Its improved understanding is of fundamental importance when looking for *solutions to manage the increased data usage and to improve the quality of communication service provided*. The resulting knowledge can help to design more adaptable networking protocols or services, as well as to determine, for instance, where to deploy networking infrastructure, how to reduce traffic congestion periods, or how to fill the gap between the capacity granted by the infrastructure technology and the traffic load generated by mobile users.

A significant amount of works in the literature analyze network traffic usage through voice calls and SMS messages, both extracted from traditional Call Detail Records (CDRs). Analysis such as [4, 9, 8, 6, 10, 11, 12] may provide an idea on the activity of mobile network customers but do not describe realistic data traffic demand patterns. In fact, contrarily to data traffic demands, call traffic has the limitation of being sparse in time (i.e., generated only when a voice call or a text message service occurs), which makes cellular users invisible at all other periods of time. Moreover, *due to the richness of the data set used in our studies, we can precisely infer traffic activity patterns over time, instead of considering only the times at which users actively generated traffic*. This includes the traffic load automatically generated by current smarphone applications (email checks, synchronization, etc). Our analysis also differs from [6, 13], since we target an individual user characterization rather than a network-wide one. Additionally, we aim to profile subscribers by their traffic demands, not by their browsing behavior, i.e., websites they normally visit as proposed in [14]. Moreover, contrarily to [8], we focus on an activity pattern characterization of a normal day, known to represent typical network usage.

Still, other works such as [7, 3, 15, 16, 17, 18], or [19] have categorized actual mobile traffic usage. For instance, [7, 3, 15, 16], and [19] have only considered total traffic volume when characterizing users' behavior. Studying this metric alone does not reflect the activity variation of users: i.e., number and frequency of requests. [17, 18] study the distribution of mobile traffic volume among different areas in a specific region. Their study, however, is based on the normalized volume with respect to the total traffic volume in the region. Instead, *we provide a precise temporal network usage characterization of a routinary day of users' life, which considers volume of traffic demand as well as its frequency. In this context, users behaviors over time are individually analyzed, with no normalization performed. Similarly, we consider activity patterns and a profiling of individual users behavior*.

With regards to age, gender and network usage investigation, [11, 12] are the most prominent works in the literature. Both studies analyze how gender and age affects the usage of voice calls and text messages, but contrarily to our work, it does not provide data traffic analysis.

3. Dataset

Our dataset captures subscribers' traffic activities generated by 6.8 million smartphone devices located within the large urban area of Mexico city. The data includes information about subscribers' *sessions* that took place from 1st July to 31st October, 2013. In the 3G standards, 3GPP or 3GPP2, a session is created when the radio channel is allocated to a subscriber as soon as he has data to

be sent. The radio channel might be seen generically as a radio resource, e.g., time slot, code, or frequency. The network finishes a session after a subscriber's period of dormancy, which is configurable and typically set from 5 to 30 seconds [20, 21]. The studied dataset contains more than 1 billion sessions and each of them has the following information fields: (1) upload and download volumes (in kilobytes); (2) session duration (in seconds); and (3) timestamp indicating when the session starts. No device-specific information is available. Furthermore, information about age and gender for 548,000 subscribers is available.

Due to the routinary behavior of people [2] and the large scale dataset, it suffices to study a subset of the large-dataset in order to capture the daily behavior of subscribers. Indeed, our analysis shows a low day-to-day variability of subscribers' activity. Therefore, we selected one week to assess the subscribers' behavior in more detail. We select the days from the 25th August to 31st August 2013 given they capture a week with typical number of devices (about 2.8 million) and activity (about 104 million sessions) (see 1(a) for a better visual comparison). This week has no special days or holidays and it is not part of the Mexican preferred vacation period spanning from early July to mid-August. From the data contained in this week, we have seen an enormous amount of outliers on the first hour of all days, likely generated by the probe when the data collection was done. Therefore, we discarded sessions starting from midnight to 1:00 am in the following analysis. This does not affect our methodology since it is indifferent to the amount of valid hours that the dataset provides. We note that while selecting a single week subset is sufficient to better assess the subscribers' behavior, we will use the whole dataset to evaluate our mobile traffic generator.

Following section, we study the behavior of mobile subscribers in terms of traffic generated. We perform the analyses according to four main traffic parameters: number of sessions, inter-arrival time (referred as IAT, the difference between the arrival timestamps of subsequent sessions of the same subscriber), volume of traffic, and session duration. Next section presents the basic notation necessary for the comprehension of those parameters.

3.1. Background Notation

Let $\mathbb{S}$ be the set of subscribers. Each subscriber $i \in \mathbb{S}$ can be effectively represented by the sequence of sessions generated by i. Let t_k^i and $t\prime_k^i$ denote the time instants at which the k-th session of subscriber i begins and finishes, respectively. $t_{k+1}^i - t_k^i$ defines then the inter-arrival time between sessions k and $k+1$ and $t\prime_k^i - t_k^i$ the time duration of session k. Let v_k^i be the volume of traffic generated by subscriber i during the k-th session and defined as the sum of the upload (i.e., $v_k^i(up)$) and download (i.e., $v_k^i(down)$) volumes generated in such session k, as provided in the dataset: i.e., $v_k^i = v_k^i(up) + v_k^i(down)$. This very fine grained representation of a subscriber is costly in terms of memory and processing time required. To overcome this drawback, we divide the studied time interval D into time slots T of duration δ. Thus, there are D/δ time slots. The notion of time slots allows us to group sessions.

For subscriber $i \in \mathbb{S}$, let τ_T^i denote the set of all sessions starting within time slot T, i.e., $\tau_T^i = \{k : (T-1) < t_k^i \leq T\}$. Now, the volume of traffic generated by subscriber i, in time slot T, is given by

$$V_T^i = \sum_{k \in \tau_T^i} v_k^i. \tag{1}$$

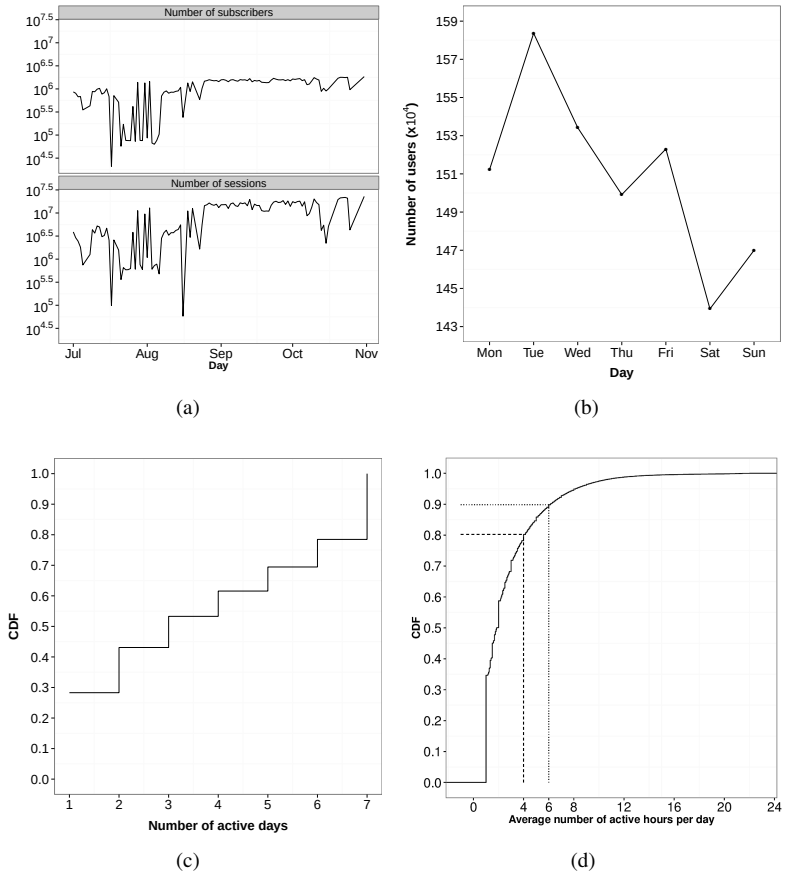
Similarly, the number of sessions generated by subscriber i in time slot T can be written as

$$N_T^i = \sum_k \mathbb{I}(k \in \tau_T^i), \tag{2}$$

where $\mathbb{I}(k \in \tau_T^i) = 1$ if $k \in \tau_T^i$; 0 otherwise. Thus, to obtain N_T^i we simply count the sessions of subscriber i that begin inside time slot T. Additionally, for each subscriber i, the average inter-arrival time in time slot T is obtained using the following expression:

$$IAT_T^i = \frac{\sum_{k \in \tau_T^i} (t_{k+1}^i - t_k^i)}{N_T^i}. \tag{3}$$

3.2. Traffic Dynamics



(a) (b)

(c) (d)

Figure 1: (a) Number of subscribers and sessions on the whole dataset. (b) Number of subscribers per day generating traffic. (c) CDF of number of days in which subscribers generate traffic. (d) CDF of number of hours in which subscribers generate traffic per day during the week.

Fig. 1(a) shows the total number of subscribers and the total number of sessions from the whole dataset. As expected, *the number of subscribers and number of sessions are highly correlated*. It is possible to see a similarity on the shape of the curves for both parameters. Indeed, Spearman's correlation between number of users and number of sessions is 98%.

In Fig. 1(b), we present the number of active subscribers, i.e., those that generated any traffic on the days throughout the selected week (recall that the selected week is 25th August to 31st August 2013). *The day-wise number of active subscribers is essentially decreasing as the week*

progresses. The difference between the weekdays and the weekend in terms of active subscribers is considerable; the highest difference is 10% which is obtained by comparing Tuesday with Saturday. As expected, *on average, the number of active subscribers are higher during the weekdays than during the weekend* (also observed in [22]). In the studied week, this average difference is 5%.

Fig. 1(c) shows the CDF (Cumulative Distribution Function) of the number of active days of the subscribers within the week (a subscriber is said to be active on a given day if she generates traffic on that day).

It is interesting to see that 22% of the subscribers generated traffic on all days, while 29% of the subscribers generated traffic only on one day of the week. Also, 53% of the subscribers generated traffic on three or less days during the week. Similar percentages were measured from a different dataset and reported in [22].

Similarly, in Fig. 1(d), we show the CDF of the average number of active hours of the subscribers per day. We see that *most of subscribers generate traffic during few hours of the day.* Indeed, on average 80% of the subscribers generate traffic for up to 4 hours each day (dashed line). If we consider a longer period, e.g., for up to 6 hours, the number of such subscribers reaches 90% (dotted line).

Fig. 2(d) shows the total number of sessions per user per day of the week. *There is a slightly smaller amount of sessions per user during weekends, and a general similarity between the cumulative values for all days.* For instance, considering users with up to 10 sessions per day, the difference between the number of sessions per user on weekdays and weekends is 4%, and 0.1% considering up to 100 sessions per day.

Fig. 2(a) presents the CDF of session duration per subscriber during the week. *We see a median usage of 63 seconds per session and a significant variation in the duration of sessions.* Interestingly, most of the sessions present a short duration, and few subscribers (less than 1%) have sessions of more than 6 hours during the week. In particular, the duration of 58% of the sessions is at most 100 seconds, while 90% of the sessions last at most 15 minutes ([22] reports similar behavior).

Fig. 2(b) shows the CDFs of the average upload and download volumes of traffic generated per session. Observe that both the upload and download CDFs are similar: e.g., 35% and 38% of the sessions, respectively, present upload and download volume of up to 1 MB. On the other hand, 6% and 13% of the sessions present more than 100 MB for uploaded and downloaded volume, respectively. *We observe that the median traffic load generated by typical subscribers is not significant, while a small number of "heavy hitters" consume a significant amount of network resources.*

Fig. 2(c) shows the *hexagonal bin plot* [23] of uploaded and downloaded volumes per session during the week. The intensity of a bin represents the frequency of sessions that generated upload and download volumes laying within the bin. *The hexagonal bin plot reveals an uphill pattern from left to right, indicating a positive monotonic relationship between the per-session uploaded and downloaded volumes.* That is, if the amount of downloaded traffic is higher in a session, we can expect the uploaded volume to be higher as well. Indeed, the Spearmans's correlation coefficient between per-session upload and download traffic is 88%. Although monotonically related, upload and download do not have high linear correlation, Pearson's correlation is 56%, probably caused by two groups of bins forming straight lines, close to each axis. Bins close to the x-axis are due to sessions that present a small upload volume, e.g., around 1 KB, and significantly higher amount of download. Those are likely sessions in which subscribers use streaming media sites that typically use Real Time Protocol (RTP). RTP does not require the subscribers' device

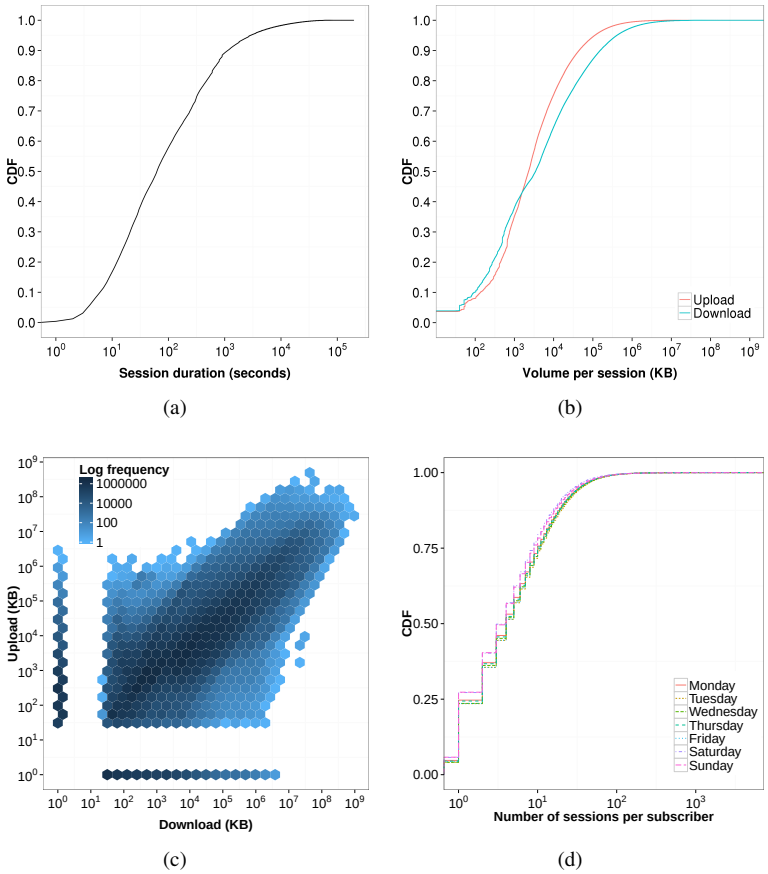


Figure 2: (a) CDF of session duration in seconds per subscriber during the week. (b) CDF and (c) bin plot of the upload and download volume during the week. (d) Number of sessions per subscriber per day of the week.

to generate confirmation packets, which justifies the small amount of uploaded volume. On the other hand, bins close to the y-axis represent sessions with small amount of download and comparably higher amount of upload. That is probably due to upload of media formats, e.g., photos on Facebook or videos on Youtube.

Owing to the high correlation between the upload and download volumes, in our evaluation and traffic modelling, we take into consideration the total volume per session, i.e., the sum of the upload and download volumes during the session.

3.3. Temporal Dynamics

It is common knowledge that some hours tend to be more active than others when it comes to users routinary daily activities. In this context, peak hours present high frequency of requests and volume of traffic, while non-peak hours present less traffic demands and volume. Indeed, Figs. 3(a), 3(b) and 3(c) show three parameters and their hourly dynamics during the week. Two features are important to highlight: *First, there is a repetitive behavior during different days at the same hours. Second, there are peak and non-peak hours when it comes to subscribers' traffic*

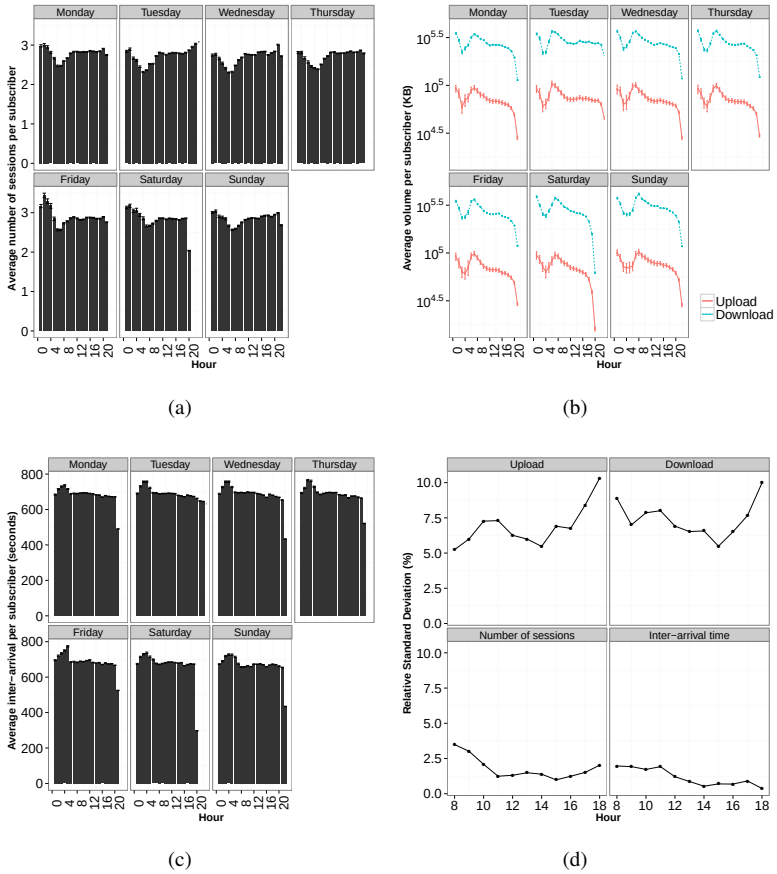


Figure 3: (a) Average number of sessions per user during the week. (b) Volume of traffic for upload and download during the week. (c) Inter-arrival time per subscriber during the week. (d) Relative Standard Deviation per parameter.

demands. In the following, we discuss these features and measure how repetitive their behavior is. We further develop the idea of peak and non-peak hours for the users' activity in our traffic model.

Fig. 3(a) shows the average number of sessions per subscriber on each hour during the studied week. The results show a clear gap on the average number of sessions from 4am to 8am. *On the end of late night and beginning of the day subscribers tend to perform less sessions*. This is consistent with diurnal human activity patterns. The number of sessions generated from 4am to 8am is 10% less when compared with that generated during the rest of the day. Furthermore, the total number of sessions from 9am to 3am is 47% higher than from 4am to 8am. Such behavior repeats over all days of the week.

Fig. 3(b) shows the upload and download session volumes per user during the week. *Similar to the number of sessions behavior (Fig. 3(a)), it is possible to see both: the gap between 4am to 8am and the day-wise similarity*. Besides, there is a small difference between the traffic on weekdays and weekends (similar behavior reported in [22]). It is challenging to precise the reason of such behavior, e.g., result of a socio-cultural aspect from Mexico city at the period of the data collection. Another possibility is the presence of Machine-to-Machine (M2M) communication in the network [24]. However, information from the mobile operator does not indicate the existence

of such data inside the studied dataset.

Fig. 3(c) shows the inter-arrival time (IAT) of subsequent sessions of the same subscriber. The high IAT shown from 4am to 8am is a complementary behavior to the low average number of sessions on the same hours present in Fig. 3(a). This is expected and due to the fact that *longer inter-arrival times results in less number of sessions on average*.

In summary, these last three results show *a high day-wise similarity on number of sessions, volume of traffic, and inter-arrival time traffic parameters. Indeed, all traffic parameters have similar per-hour values on different days, even comparing weekdays and weekends*. We measure the day-wise variability on subscribers' behavior using the Relative Standard Deviation (RSD). RSD is the absolute value of the coefficient of variation (CV), which is the ratio of the standard deviation σ to the mean μ. Fig. 3(d) shows the per-parameter average RSD, which considers the hour-wise variation from all 7 days during Mexican working hours (i.e., from 8am to 6pm). It is calculated using the values of the parameters of the same hours for all the days, e.g., the RSD for the number of sessions at 10 a.m. among all days is 2.08%. The average variability is small for all parameters: 1% for number of sessions, 1% for inter-arrival time, 6% and 7% for upload and download volumes, respectively.

In order to show that the variability within the day is higher than the variability among the same hours on different days, we have calculated the maximum RSD of each parameter on all hours of each day. The results are, on average, 4% for number of sessions, 2% for inter-arrival time, 16% and 15% for upload and download volumes, respectively. Therefore, we can conclude that, *on the studied dataset, the parameters from the same hours on different days present less variability than the parameters within the same day on different hours*. Note that this conclusion cannot be attributed to the variation in the number of users during the week, since it does not change significantly as shown in Fig. 1(b).

Contrarily to our findings, previous related studies considering phone records (or CDRs) [3] show that subscribers behavior in terms of call traffic have strong variation with time and day of the week. Instead, our results show the consideration of real data traffic (instead of call traffic) (1) reveals a different facet of subscribers behavior and (2) stresses the imprecisions brought by CDRs analysis to the resource allocation planning.

The similarity of the temporal activity patterns among different days of the week is due to people's natural routinary behavior. Therefore, *we select one day (28th August 2013, a Wednesday) of the week to perform our extensive per-hour analysis and distinguish users profiles.* Although we have selected this precise day, we use the whole dataset to evaluate our final traffic generator model.

3.4. Age and Gender Dynamics

Gender and age analysis are of strong interest of network operators. As discussed in the references [11, 25], gender and age analysis provide knowledge on the structure and demographics of the mobile phone market. In fact, a better description of this ecosystem can give the mobile operators a business edge in areas such as churn prediction, targeted marketing (e.g., specific health campaigns for women), and better client service among other benefits. Among the 2.8 million subscribers in the week mentioned in Section 3.2, a subset of 548 thousand users present personal information regarding age and gender. The analyses in this section refer to this subset of users. Thus, to better understand how age and gender impacts traffic demands, we present hereafter our analysis on the traffic parameters when considering this new social information.

As any study considering social aspects of participating entities, it is important to understand in which cultural context the measurements are made. Like many Latin American countries,

Mexican society presents a gender wage gap that disfavours women [26]. Consequently, having less purchasing power, Mexican women consume less goods. As a probable consequence, from almost half million users of the considered dataset, 56% are men and 44% are women.

Fig. 4(a) depicts the population histogram grouped by age and gender. This graphic shows the frequency of age and genders' occurrences with females on the left and males on the right. *Regardless of the gender, it is possible to see a higher number of subscribers with age range from 25 to 34 years old.* Indeed, 33% of the subscribers fall in this range.

To ease the understanding of the per-age behavior, we have defined 4 age ranges as in [25]: $[15, 24]$, $[25, 34]$, $[35, 49]$ and $[50, 85]$, i.e., users younger than 25, from 25 to 34 years old, from 35 to 49 years old, and older than 50 years old. We have removed users younger than 15 and older than 85 years old from the trace: the small amount of users in those two groups make it difficult to draw any statistical conclusion about them. Fig. 4(b) shows the percentage of subscribers grouped by gender and age ranges. It is possible to see a higher percentual of male (and consequently less female) users in all age ranges. An interesting aspect of this graphic is the increasing gap between the genders as the age range progresses. To undercover this aspect, we have plotted Fig. 4(c). It shows the percentage of users per age and gender. It is interesting to see that the gap augments as the age increases. The Spearman's correlation between age and age percentage per gender is 87% per male and, consequently, -87% per female, i.e., in our dataset the *male participation percentually increases as the user age increases. Conversely, the female participation decreases with the increase of the age.*

Fig. 4(d) shows the percentage of active users per age and day of week. An interesting aspect in this graphic is on Saturday and Sunday. These two days have different age range activities when compared to the rest of the days of the week: in particular, we observe the absence of the gap present on weekdays from 4 am to 8 am for users within the $[25, 34]$ range, indicating an activity growth for users within this range. This is probably due to the nightly activities that usually attracts younger people on weekends, e.g., bars and night clubs.

Fig. 5(a) shows the frequency of the number of sessions by subscribers grouped by their ages per day of the week. For readability reasons, the plot shows subscribers with up to 500 sessions. Still, the figure depicts 99.99% of the data related to subscribers' number of sessions. Like the day-wise similarity presented in Section 3.3, this graphic shows that the age-wise number of sessions is similar on different days of the studied week. Regardless of the age, most users present low and similar number of sessions per day (see Fig. 2(d)). Briefly, *per age behavior shows that younger subscribers tend to have peak number of sessions that are higher than older subscribers.*

Fig. 5(b) better shows the decreasing behavior of the traffic parameters as the age increases, regardless of the gender. It depicts the mean of four traffic parameters by user, grouped per age and gender. As the number of users older than 70 years old is small, their mean values tend to be noisy. If we consider users up to 70 years old, there is a high negative correlation between age and each of the traffic parameters, for males and females: respectively -96% and -95% for volume of traffic, -85% and -71% for number of sessions and -63% and -78% for session duration. It means that as the age grows, the value of each of those traffic parameters decreases. Except for the inter-arrival time, there is a clear gap between the maximum and the minimum values for each of the parameters from younger to older subscribers, in particular for the total volume of traffic. In order to measure this difference, we have calculated the fraction of the traffic parameters from the oldest age range divided by the youngest one. Indeed, *users from the youngest age range generate, on average, 52% more traffic volume, 21% more sessions, 12% longer sessions with the same inter-arrival time.* In our dataset *users' network activity tend to decrease with the increase of their age.* Our analysis also show the same decreasing activity when subscribers are grouped by their

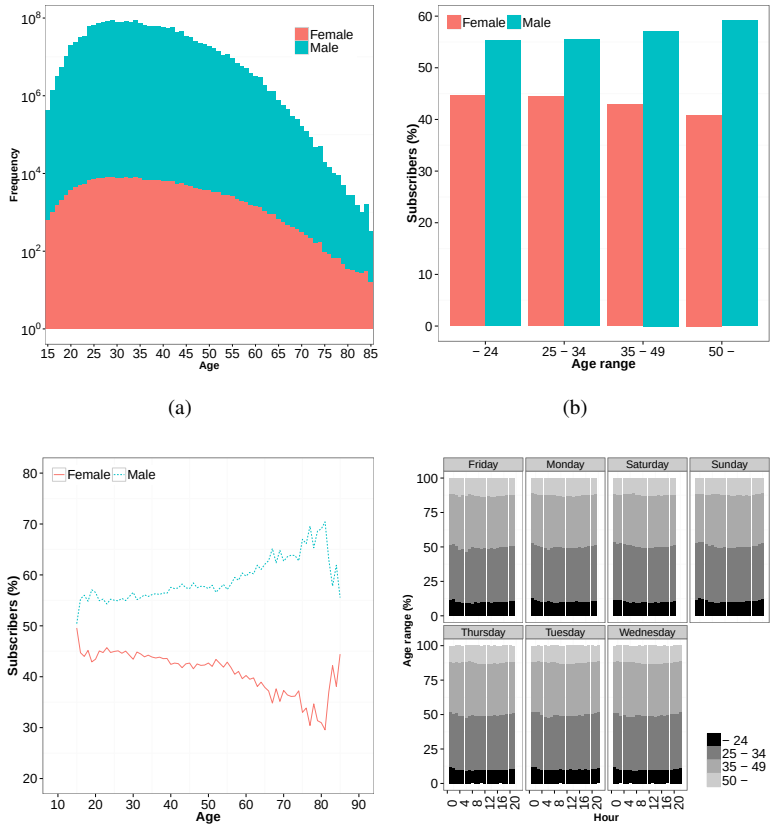


Figure 4: (a) Population pyramid grouped by age and gender. (b) Subscribers by gender per age ranges. (c) Percentage of active users by age. (d) Percentage of active users by age range.

genders, i.e., *it is related to the age of the subscribers and not a behavior of a specific gender*.

Fig. 5(c) and 5(d) show the CDF of number of sessions and CDF of session duration, respectively, grouped by age range and subscribers' gender. As already discussed, the *mean network demand is higher for younger users than for older users.* Grouping users by age range diminishes this gap when compared to the per-age analysis, but still allows us to see the cumulative differences. For both genders: (1) 80% of the subscribers of the oldest age range and 76% of the youngest age range generate up to 10 sessions during the day, and (2) 48% of the subscribers of the oldest age range and 43% of the youngest age range generate sessions of up to 15 minutes during the day. *In summary, our analysis show that similar number of sessions and session duration results are seen when users are grouped by age range, irrespective of the subscribers gender.*

4. Subscriber Profiling Methodology

CDR datasets typically include information about millions of users, collected during several months. The common practice, when considering data traffic, is to process the whole data at

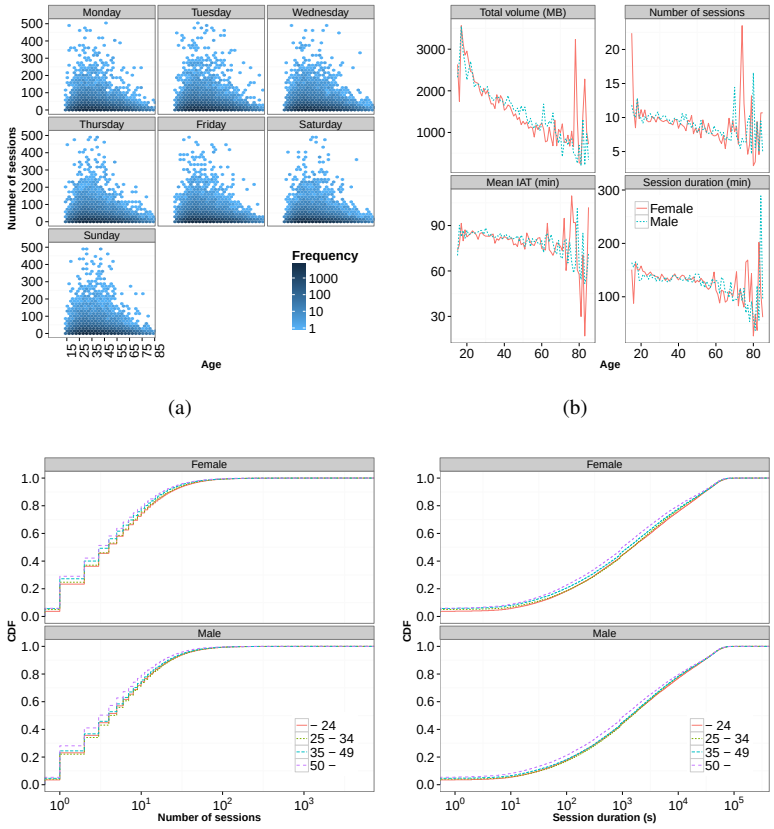


Figure 5: (a) Frequency of sessions per age and day. (b) Mean metrics per age and gender. (c) CDF of the number of sessions per age range and gender. (d) CDF of the session duration per age and gender.

once [22], possibly identifying geographical or temporal patterns. Nevertheless, indiscriminate aggregation of large CDR datasets may delineate global trends, but it completely loses information on outlying user behaviors. In fact, although having their own repetitive routine, human behavior in terms of content demand is highly heterogeneous, as many other human activities. While some subscribers rarely generate mobile data traffic, others demand a few or even a large amount of gigabytes each day (i.e., in an aggregated analysis, heavy hitters would dominate the traffic results and hide the modeling of light users' traffic).

To analyse such different levels of activity, we group subscribers into a limited number of profiles from a training set and classify network usages accordingly. In particular, the profiles are defined according to two traffic parameters: traffic demands (i.e., volume of traffic) and activity behavior (i.e., number of sessions). Such parameters are extracted from a subset of the considered dataset. The profile definition is performed in three phases. First, the similarity metric between all pairs of subscribers is measured according to the traffic parameters. Second, subscribers are clustered by their similarity into a limited number of clusters, which represent user profiles. The third phase consists in classifying the remaining subscribers of the dataset into the previously defined profiles. This profiling procedure results in typologies of subscribers based on their traffic

dynamics. The remainder of this section details these different phases.

4.1. Similarity Computation

Although we later evaluate our methodologies for a day within the week, our development in this section can hold in general for any *time interval* D chosen from the week; formally, D is a collection, $\{T\}$, of successive time slots (recall Section 3.1). For a given time interval D, let $\mathbb{S}$ be the set of all subscribers that generate some traffic during D, and $\mathbb{S}' \subseteq \mathbb{S}$ be a randomly selected sample of subscribers from $\mathbb{S}$. Our objective is to partition the subscribers in $\mathbb{S}'$ into a set of *clusters* $\mathbb{P}$, such that subscribers belonging to the same cluster are "similar" in terms of traffic demands. We use Euclidean distance to measure the *similarity* between two subscribers [27]. We then *classify* the remaining users in $\mathbb{S}$ (i.e., $\mathbb{S} - \mathbb{S}'$) into various clusters in $\mathbb{P}$. In this work, we develop a similarity comparison in terms of *volume of traffic* and *number of sessions* for time interval D (as defined hereafter). These traffic parameters allow us to compare different subscriber behaviors, and will be considered during the clustering and classification procedures (discussed in the next section).

Using Eq. (1) and Eq. (2) from Section 3.1, the total volume and the total number of sessions generated by subscriber i during D are computed as follows: $\vartheta^i = \sum_{T \in D} V_T^i$ and $\eta^i = \sum_{T \in D} N_T^i$.

Finally, we define the *traffic volume similarity* between two subscribers i and j as the difference between the total volumes generated by these users, i.e.,

$$w_{ij}^{\vartheta} = \|\vartheta^i - \vartheta^j\|. \tag{4}$$

Likewise, the *number of sessions similarity* can be defined as

$$w_{ij}^{\eta} = \|\eta^i - \eta^j\|. \tag{5}$$

Using the subscribers in $\mathbb{S}'$ as the vertices, and using either $w_{i,j}^{\vartheta}$ or $w_{i,j}^{\eta}$ as the edge weights, we obtain a complete graph $G(\mathbb{S}', \mathbb{E})$, which is given as input to our clustering algorithm to obtain the set of clusters $\mathbb{P}$. The remaining users (i.e., $\mathbb{S} - \mathbb{S}'$) are then classified into the previous defined clusters.

4.2. Subscriber Clustering and Classification

Instead of a-priori fixing a value for the number of profiles (i.e., clusters) $|\mathbb{P}|$, our goal is to obtain from the data the number of profiles which best represent the subscribers' traffic activities. For this purpose, we use an hierarchical clustering algorithm that iteratively aggregates vertices from the similarity graph $G(\mathbb{S}', \mathbb{E})$ into larger clusters, according to a dendogram structure [28]. The hierarchical clustering algorithm we choose is the *Average Linkage clustering method*, also known as *Unweighted Pair Group Method with Arithmetic Mean (UPGMA)* [28].

Recall we first group a sample set of $|\mathbb{S}'|$ subscribers into $|\mathbb{P}|$ clusters, and then we classify the remaining $|\mathbb{S} - \mathbb{S}'|$ subscribers into $\mathbb{P}$. Thus, UPGMA starts by first considering each vertex of the given graph $G(\mathbb{S}', \mathbb{E})$ as a cluster (i.e., singleton clusters). At each iteration, it computes the distance (using the edge weights between vertices given by Eq. (4) or Eq. (5)) between all pairs of clusters, and then merges the closest two clusters. In our context, it merges together the two clusters that are more similar in terms of traffic demands. If the algorithm is not stopped, it finally simply yields a single cluster containing all the vertices.

Thus, it is important to find where UPGMA should stop its merging process, yielding the best number of clusters, i.e., *the best separation among the groups of usage pattern from subscribers*.

To that end, we use several *stopping rules* (or stopping criteria). A stopping rule, during each iteration of the hierarchical clustering algorithm (or each level of the dendogram), gives a measure of how well separated the clusters are, based on which one can decide the best number of clusters to use.

In the literature, there are several stopping rules [29]. Contrarily to related work that have implemented and applied very few of them [22] and in order to avoid to be biased by a specific criteria, we have implemented and used 23 stopping rules. For a complete list of the implemented stopping rules, please refer to [30].

For the sake of illustration, we briefly describe the C-Index stopping rule. C-Index is defined as $C = (S - S_{min})/(S_{max} - S_{min})$, where: (1) S is the sum of all distances between pairs of users in the same cluster over all clusters, (2) S_{min} and S_{max} are the sum of the smallest and the largest distances respectively, for all pairs of users, over all clusters. In our context, it compares the distances among the considered traffic parameters. According to C-Index, the lower the value of the index, the better the clustering. In this way, the number of profiles producing the lowest C-Index value is the one that grants the best separation among clusters.

Fig. 6(a) shows the C-Index index values as a function of the number of clusters, when the number of sessions similarity is considered as the distance between pairs of users. The best number of clusters is given by the minimum C-Index value. Thus, the best number of clusters is 2 according to Fig. 6(a).

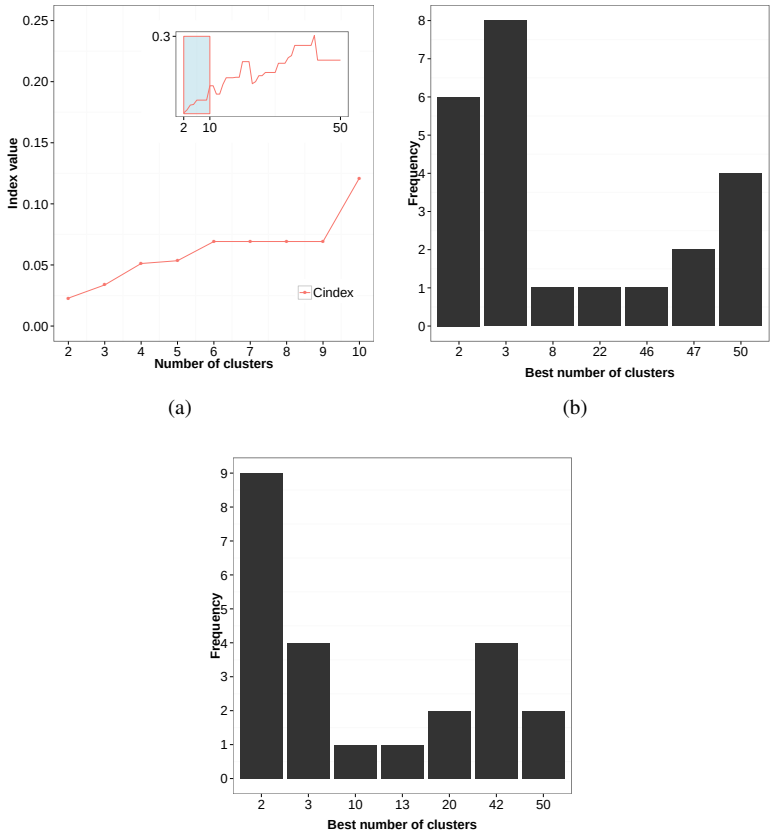
Similarly, each of the 22 other stopping rules implemented define their best number of clusters. In Fig. 6(b), we present the frequency of the best number of clusters, while profiling subscribers using traffic volume similarity. It condensates in a histogram the result of the 23 stopping rules. It shows that 8 stopping rules recommend 3 as the best profiles, when clustering subscribers by their traffic volumes.

Summarizing, profiling occurs in four stages: (1) building a similarity graph with $|\mathbb{S}'|$ subscribers, (2) hierarchically clustering it using a similarity metric, (3) determining the best number of clusters $|\mathbb{P}|$, by relying on the stopping rules, and (4) classifying $|\mathbb{S} - \mathbb{S}'|$ remaining unclassified subscribers in the previously defined clusters.

In the fourth stage, we use the *k-means algorithm* as the classification technique. It is worth mentioning, we calculate the clusters centroids (means) obtained from the hierarchical clusters and use them on the first iteration of the k-means algorithm. This is an important information because the centroids obtained from the hierarchical clustering algorithm are likely to be better positioned than the k-means originally bootstrapped initial centroids, which are based on randomly selected positions.

These four stages are performed in two rounds. In the first round, the graph $G(\mathbb{S}', \mathbb{E})$ weighted according to the *traffic volume similarity* (Eq. (4)) is used for the hierarchical clustering. The best number of "traffic volume"-based clusters is then determined: according to the results shown in Fig. 6(b), $|\mathbb{P}| = 3$ weighted subgraphs $\{G_1(\mathbb{S}'_1, \mathbb{E}), G_2(\mathbb{S}'_2, \mathbb{E})\}, G_3(\mathbb{S}'_3, \mathbb{E})\}$ are created. At the end of the first round, the final classification of $|\mathbb{S} - \mathbb{S}'|$ subscribers takes place.

The next execution round initiates with a new hierarchical clustering being performed inside each initially defined "traffic volume"-based cluster. This time G_1, G_2 and G_3 are weighted according to the *number of sessions similarity* (Eq. (5)). Finally, for each of these three initial clusters, two "number of sessions"-based clusters are defined after the second round of stopping rules execution (e.g., Fig. 6(c)), totalizing six subscribers profiles. Due to space constraints, we will not show all stopping rules results. The second round ends with the classification of the remaining $|\mathbb{S} - \mathbb{S}'|$ subscribers into the six defined profiles. Next section better details our subscriber profiling.



(a) (b)

(c)

Figure 6: (a) C-Index values and respective number of clusters when re-clustering subscribers at the 3rd defined "traffic-volume"-based cluster, according to the number of sessions similarity. (b) Histogram of best number of "traffic-volume"-based clusters indicated by the assessed stopping rules. (c) Histogram of the best number of "number of sessions"-based clusters indicated, when re-clustering subscribers using the second "traffic-volume"-based cluster.

4.3. Subscriber Profiles

To obtain the profiles for our dataset, we set D as 28th of August, which contains information of about 1.5 million smartphone devices, and randomly sample 10000 subscribers (thus, $|\mathbb{S}'| = 10000$ to be used in the clustering procedure, which is the maximum number of elements allowed in the hierarchical clustering package of the R language.). D is a normal day in the middle of the week (i.e., likely less influenced from weekend-related behavior) with no special event or holiday and we divide it into time slots of duration δ. Time slots help to understand the general behavior of a certain period of time in D. The higher the number of time slots, the shorter their duration, and vice-versa. Very short time slots, e.g., 1 minute, may lead to an analysis with fewer sessions per time slot, hindering the identification of subscribers' behavior per slot. Very large time slots, e.g., 12 hours, may lead to a general view of the sessions, which make it difficult to obtain a good quality assessment of the traffic dynamics. Thus, for our evaluation, we choose a "moderate" value of $\delta = 1$ hour as the time slot duration. Nevertheless, the optimal size of the time slot is still an open problem [31].

Our profiling methodology resulted in *six profiles*, and we have named them as follows: Light

Occasional (LO), Light Frequent (LF), Medium Occasional (MO), Medium Frequent (MF), Heavy Occasional (HO) and Heavy Frequent (HF). *Light* profiles contain subscribers that generate up to 17 MB of data during the day, *Medium* profiles have subscribers that generate between 17 MB and 560 MB of traffic during the day, and *Heavy* profiles contain users that generate more than 560 MB of traffic during the day. Likewise, *Occasional* profiles contain subscribers that generate less connection sessions, whereas *Frequent* profiles contain users generating more connections per day. Table 1 shows the characteristics of each of the profiles.

Table 1: Characteristics of the resulting profiles

	Light		Medium		Heavy	
Volume	29 KB to ≈ 17 MB		17306 KB to ≈ 560 MB		560046 KB to ≈ 650 GB	
Nº of subscribers	418843		610917		487141	
	Occasional	Frequent	Occasional	Frequent	Occasional	Frequent
Nº of sessions	1 to 10	11 to 224	1 to 51	52 to 1926	1 to 316	317 to 8737
Nº of subscribers	405848	12995	598340	12577	484959	2182

In Fig. 7, we show the dynamics of the traffic parameters per subscribers' class per hour. Fig. 7(a), 7(b), and 7(c) corresponds to the number of sessions, volume of traffic, and mean inter-arrival time, respectively; the error bars correspond to a 95% confidence interval. For each time slot, the volume of traffic, number of sessions, and inter-arrival time are calculated using Eq. (1), Eq. (2), and Eq. (3), respectively.

From Fig. 7, we can see that our methodology well separates the profiles, i.e., the *occasional* and *frequent* subscribers have their values clearly separated. *Note that an aggregated traffic analysis would not allow us to identify and consequently, to imitate the behavior of very light users. In fact, the traffic generated by very heavy users (representing a very small percentage of users in the dataset) would bias the analysis and the synthetic traffic generation.*

For each curve in Fig. 7(a), 7(b), and 7(d), we have also shown a *horizontal line that represents the respective mean value* (where the mean is taken over all time slots). Given the mean values, we classify, for each profile of subscribers and for each parameter (number of sessions, traffic volume, and IAT), the hours above the mean as *peak hours*, and hours below the mean as *non-peak hours*.

4.4. Profile's Age and Gender

In this section, we assess each of the resulting profiles by the age and gender of their members. The profiled day D has 1.5 million users, from which 107 thousand have information regarding age and gender. The results shown in this section refer to this subset that counts with 57.6% of male and 42.4% of female users. This subset is consistent with the distribution of users with available age and gender prior to the profiling process, which counted 548 thousand subscribers over a week (Section 3.4). To evaluate this consistency, we calculate the percentage of users per age on the 548 thousand non profiled users and on the 107 thousand profiled users. Fig 8(a) shows this percentage for each of them. There is a visual similarity between the shape of the two curves as they are strongly correlated, with 99% Spearman's correlation.

Fig. 8(b) shows the percentage of male and female subscribers per class, after the profiling of 107 thousand subscribers. *Most of the classes present higher percentual of male than female, except HF in which female have 1% more users than male.* On average, Light and Medium profiles have 15% more males than females, while Heavy profiles have 6% more male than female.

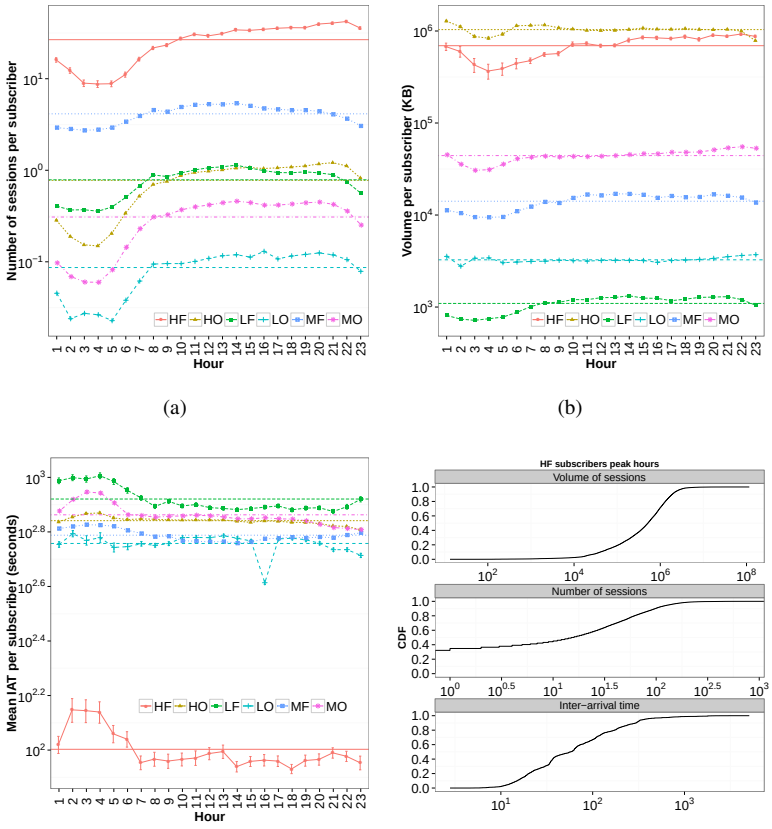


Figure 7: (a) Number of sessions per class. (b) Volume of traffic per class. (c) Mean inter-arrival per class. (d) Empirical CDFs of HF users in peak hours.

Fig. 8(c) shows the average subscribers' ages per gender and classes. Due to the large overlapping presented by the confidence intervals (95%), we can assert that *the per-class ages are not significantly different*. That is interesting because *it indicates that the profiles group together users from a wide spectrum of different ages.*

Fig. 9(a) and 9(c) show the CDFs of number of sessions per subscriber per class. The former groups subscribers per age range and the latter per gender. An interesting difference between Occasional and Frequent users is the steepness of the CDF curves. *The number of sessions for Occasional profiles is more uniformly distributed than for Frequent users, which has a very steep slope. It means that most of the Frequent users generate the lowest amount of sessions within the range of their profiles* (recall that Table 1 specifies the ranges). *For all classes, male users generate, on average and median, more sessions than females.* On Occasional classes the difference is 1% at most, while on Frequent classes the difference ranges from 2% to 19%. The cumulative values show the same results, for instance the third quartile is at most 1% higher for male than female on all Occasional and LF profiles. Moreover, it is 10% higher on MF and HF profiles.

Fig. 9(b) depicts the CDFs of session duration per subscribers' class and age range. On

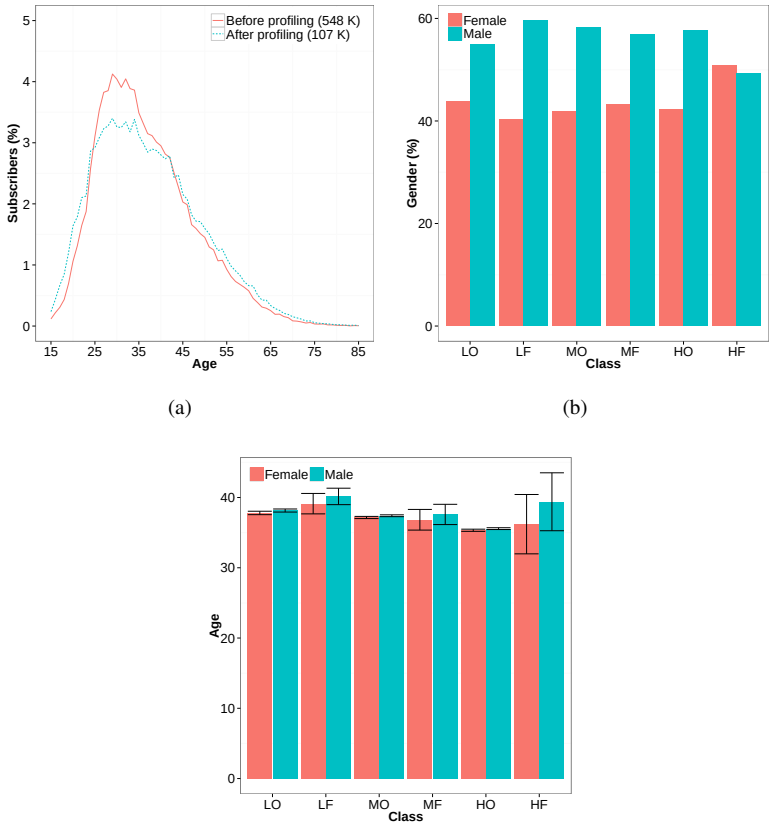


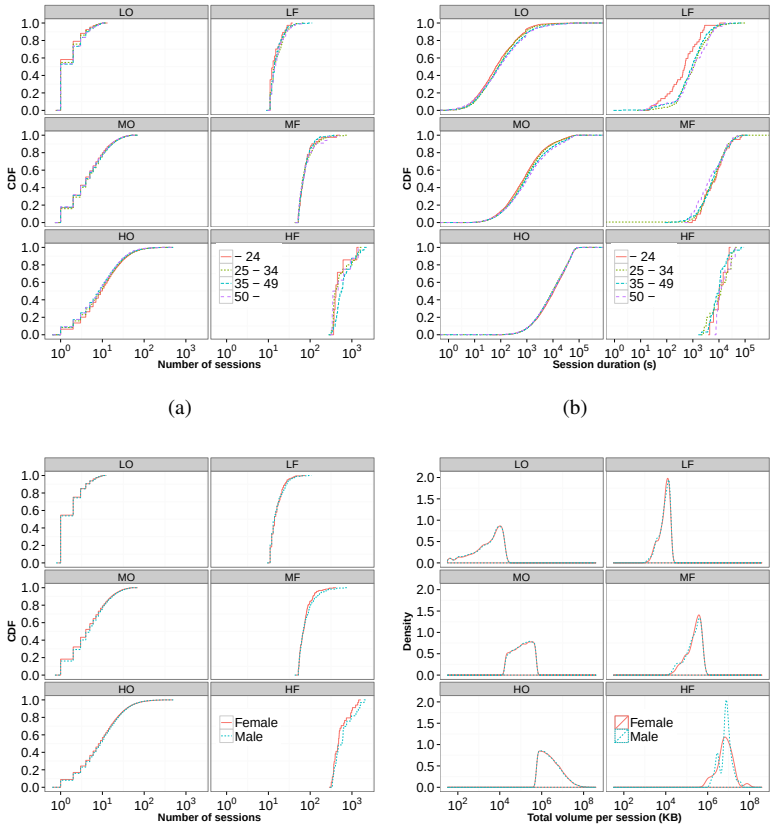
Figure 8: (a) Percentage of subscribers per age before and after profiling. (b) Percentage of subscribers per gender and class. (c) Average subscribers' age per gender and class.

average, profiles do not present statistically different session duration values for each of the age ranges. For instance, the per-class confidence intervals (95%) for each of the age ranges overlap each other by the mean. It means, *the session duration behavior within each of the profiles for a certain age range is not statistically different from the behavior of another age within the same profile.*

Fig. 9(d) presents the kernel density estimation (KDE) curves for the volume per user per gender and class. *There is a similar behavior for male and female subscribers for all the profiles, except HF.* HF male subscribers density curve is narrower than the female one and present a peak around 10 GB. On the other hand, HF female subscribers curve is wider. It means that, *among the heavy and frequent subscribers, male present less diverse session volumes when compared to female.*

5. Measurement-driven traffic modeling

Realistic network simulations require a traffic generator capable of imitating actual daily subscribers' traffic demands, i.e., has to be consistent with the observations made about the real



(a) (b)

(c) (d)

Figure 9: (a) CDFs of number of sessions and (b) session duration per subscribers' class and age range. (c) CDFs of number of sessions and (b) session volume per subscribers' class and gender.

subscribers in the previous section. Recall that subscribers belonging to different profiles (LO, LF, MO, MF, HO, and HF) have their own specificities in terms of *when* the sessions are generated during the day, and the *volume* generated during each session. Furthermore, each profile of subscribers have different behavior during *peak* and *non-peak hours*. Thus, to obtain a fine grained model it is important to take into account all the above considerations, while generating a synthetic trace. Therefore, we model profile of subscribers described by their traffic parameters distribution in each type of hours.

The traffic modeling is thus performed in two main phases. First, for each subscriber profile, we estimate the set of distributions that better fit the distribution functions of each traffic parameter (i.e., traffic volume, number of sessions, and inter-arrival time) during peak and non-peak hours. For this, adapted statistical tests are selected according to the type of values of each parameter, i.e., continuous or discrete values. Second, once the list of best fitted distribution along with their parameters are determined, a subscriber is synthetically generated: (1) A profile is randomly assigned to each synthetic subscribers according to the distribution of profiles population in the real dataset and (2) according to the assigned profile and for each daily hour, values for each traffic parameter are randomly sampled from he corresponding fitted distribution functions. In the

following, we describe how we merge all the above considerations to obtain a measurement-driven mobile data traffic modeling.

5.1. Fitting Empirical Distributions

Using the original subscribers' data, we first study for each profile in peak and non-peak hour, the empirical distribution functions (i.e., CDF) of the traffic parameters: the number of sessions generated, the traffic volume associated with each of these sessions, and the inter-arrival times between the sessions (Fig. 7(d) shows an example). For instance, the empirical distribution function of *"total volume for HF users in peak hours"* is obtained from the set of all V_T^i (Eq. (1)) such that $i \in \mathbb{S}$ is an HF subscriber and T is a peak hour. The empirical distribution functions of the number of sessions and the inter-arrival time for any combination of profile and hour-type (peak or non-peak), can be similarly generated using N_T^i (Eq. (2)) and IAT_T^i (Eq. (3)), respectively.

Once we obtain the CDFs, using statistical tests, we estimate the set of distributions that best fit them. From this set, we then select the closest distribution function to the respective CDF. *This function will be used at the traffic usage pattern generation for the corresponding profile and type of hour*. More specifically, when considering the volume of traffic and the inter-arrival time parameters (i.e., consisting of continuous values) of a certain profile and hour, the Kolmogorov-Smirnov statistic test [32] is used. The test estimates the parameters for a set of continuous distributions (namely, Log-normal, Gamma, Weibull, Logis, and Exponential) that best fit the corresponding empirical distribution function. Similarly, when considering the number of sessions parameter (i.e., consisting of discrete values) of a certain profile and hour, the Chi-squared statistic test [33] is used in conjunction with the Method of Moments (MME) and the Method of Maximum Likelihood (MLE) to estimate the best fitting parameters for a set of discrete distributions (Negative binomial, Geometric, and Poisson). In both cases, after obtaining the results from the fitting test, we select the distribution functions that best fit each corresponding CDF.

Table 2 list the best fitted distribution functions along with their parameters for all possible combinations of profile and hour-type pair for number of sessions. Refer to [30] for the complete list of tables, including traffic volume and inter-arrival time. For the Negative-binomial distribution, n is the size parameter and p is the probability parameter.

5.2. Synthetic Subscriber Generation

Generating a synthetic subscriber will first require us to generate a profile type (LO, LF, MO, MF, HO, or HF) for the subscriber. Profile types are randomly assigned, based on the distribution of profiles population observed in the real data. For instance, from Table 1, we see that 26.7% of the subscribers belong to LO profile, and thus with probability $q_{LO} = 0.267$, we assign LO profile to a synthetic user. Similarly, the probabilities of other profiles are: $q_{LF} = 0.0085$, $q_{MO} = 0.394$, $q_{MF} = 0.0082$, $q_{HO} = 0.319$, and $q_{HF} = 0.001$. We will refer to $q = (q_{LO}, q_{LF}, q_{MO}, q_{MF}, q_{HO}, q_{HF})$ as the *profile pmf*, or probability mass function.

We now briefly describe our procedure for generating a synthetic subscriber (for a detailed algorithm, refer to [30]). *We first randomly generate a profile type for a subscriber i using the profile pmf q. After obtaining the profile type, for a given hour T, we randomly sample values for each traffic parameter according to the corresponding fitted distribution functions*.

In more detail, the algorithm takes as parameter the number of synthetic users to be generated. The result of the generation is a list of sessions per user. Each synthetic user session contains two fields: (1) volume of traffic and (2) arrival timestamp. For each subscriber i and time slot T, we sample a number of sessions N_T^i (from the distribution listed in Table 2), an average session

volume V_T^i, a mean inter-arrival time IAT_T^i from the appropriate distributions (i.e., the fitted distribution corresponding to the profile and hour-type pair). The volume per session v_k^i (for $k \in \tau_T^i$, see Section 5) is then equal to the sampled value V_T^i divided by the sampled number of sessions N_T^i. The initial timestamp of each session in hour T is then computed according to the sampled inter-arrival time IAT_T^i and number of session N_T^i for that hour. By varying T over the 24 hours in a day, we obtain a synthetic subscriber traffic for one day.

Table 2: Number of sessions: distributions and parameters

Number of sessions			
Hour	*Profile*	*Distribution*	*Parameters*
Peak	HO	Negative binomial	$n = 0.1139$, $p = 0.09$
	HF		$n = 0.4703$, $p = 0.01$
	MO		$n = 0.1772$, $p = 0.3$
	MF		$n = 0.7588$, $p = 0.13$
	LO		$n = 0.1885$, $p = 0.62$
	LF		$n = 0.4802$, $p = 0.32$
Non-Peak	HO	Negative binomial	$n = 0.0448$, $p = 0.1$
	HF		$n = 0.1437$, $p = 0.01$
	MO		$n = 0.0536$, $p = 0.3$
	MF		$n = 0.3146$, $p = 0.08$
	LO		$n = 0.0810$, $p = 0.66$
	LF		$n = 0.2405$, $p = 0.33$

5.3. Synthetic Traffic Model Evaluation

In order to evaluate our traffic modeling, we generate a synthetic dataset and compare it with the original dataset. Note that, since we model profile traffic distributions and not absolute traffic values of subscribers' individual behavior, our evaluation investigates the similarity of distributions given by synthetic and real datasets. Towards this goal, we first generate a set $\mathbb{R}$ of synthetic subscribers, where $|\mathbb{R}| = |\mathbb{S}|$, and one day of traffic denoted as D'. The synthetic dataset contains for each session of subscriber i and hour T: (1) the volume in kilobytes and (2) the initial timestamp of the session.

Let $\mathbb{D}$ denote the set of days contained in the whole dataset, from 1st July to 31st October. Let p_E^{ϑ} denote the PDF (Probability Distribution Function) of the total volume generated per subscriber active in day E in the original trace, formally defined as $p_E^{\vartheta}(x) = \sum_{i \in E} \mathbb{I}(\vartheta^i = x)/|\{i \in E\}|$.

For a visual comparison, Fig. 10(a) depicts the CDFs corresponding to the PDFs p_D^{ϑ} and $p_{D'}^{\vartheta}$ of traffic generated in the original day D and synthetic day D'. We can observe an *almost complete overlap of the two CDFs due to high similarity between the real trace and the synthetic trace*.

We then assess the consistency of the synthetic traffic by comparing the distributions of the various parameters between the original and the synthetic datasets. For this, we use the Bhattacharyya (BH) measure [34]. It quantifies the similarity between two discrete or continuous probability distributions. Let $p(i)$ and $p'(i)$ be two pmfs, i.e., $\sum_{i=1}^{N} p(i) = \sum_{i=1}^{N} p'(i) = 1$. The BH measure is formally defined as $\rho(p, p') = \sum_{i=1}^{N} \sqrt{p(i)p'(i)}$. However, the BH measure is not a distance metric since it does not satisfy all the metric axioms. Therefore, [35] proposes an alternative distance metric based on the BH measure which is formally defined as $d(p, p') =$

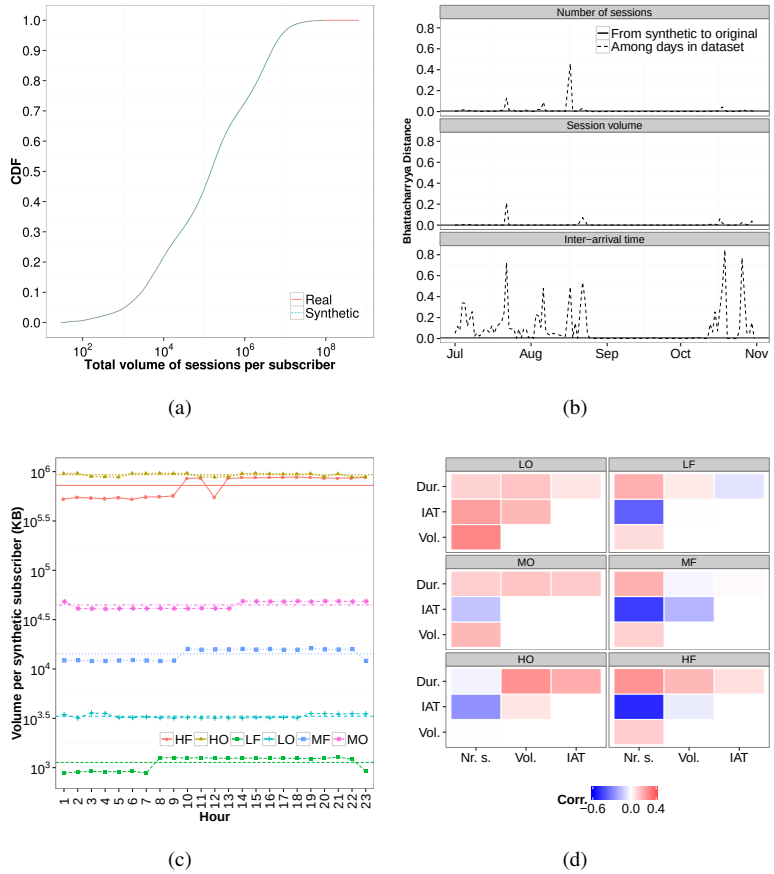


Figure 10: (a) CDF of the total volume generated by real and synthetic subscribers (b) Per-parameter BH distances between original day D and synthetic trace D' (solid line), and between the original day D and other days $E \in \mathbb{D}^*$ from the original trace (dashed line) (c) Volume of traffic per class for synthetic subscribers. (d) Heatmap (better seen in colors) of the correlation between session duration, inter-arrival time and volume of traffic.

$\sqrt{1-\rho(p,p')}$. Note that $d(p,p')$ exists for all discrete distributions, and it is equal to zero if and only if $p = p'$. We use d in order to measure the similarity between the original and the synthetic profile distributions.

Literature provides other metrics to calculate the distance between two probability distributions, e.g., Kullback-leibler (KL) divergence [36], which is well-known in the Information Theory field. In our context, the KL divergence is not suitable, since it cannot be calculated on vectors of different sizes. Since we compare the traffic demands among different days, it is unlikely that all the days contain the same number of subscribers. On the contrary, BH works regardless of the sizes of the vectors that form the CDFs and that is the main motivation on using BH instead of KL in our scenario.

Let $\mathbb{D}^*$ denote the set of days contained in the dataset, except the original day D. We first compute $d(p_D^{\vartheta}, p_{D'}^{\vartheta})$, the distance between the total volume distribution of the original day and the synthetic day. Then, we compute $d(p_D^{\vartheta}, p_E^{\vartheta})$ for $E \in \mathbb{D}^*$, the distance between the original day and remaining days in the original trace. We obtain similar distances for p_E^{η} and p_E^{ζ} (for $E \in \mathbb{D}$), which are, respectively, the PDFs of the total number of sessions and average inter-arrival time

per subscriber. In Fig. 10(b), we show as dashed lines the distances $d(p_D^{\vartheta}, p_E^{\vartheta})$, $d(p_D^{\eta}, p_E^{\eta})$ and $d(p_D^{\zeta}, p_E^{\zeta})$. Fig. 10(b) also shows, as horizontal solid lines, the distances between the original and the synthetic day $d(p_D^{\vartheta}, p_{D'}^{\vartheta})$, $d(p_D^{\eta}, p_{D'}^{\eta})$ and $d(p_D^{\zeta}, p_{D'}^{\zeta})$. Finally, for each distribution, we have also computed the mean and the confidence interval (95%) of the distances between the original day and the remaining days. The traffic model evaluation consists then in verifying whether the distance $d(p_D^*, p_{D'}^*)$ is within the confidence interval of the distances $d(p_D^*, p_E^*)$. Although not shown here, we verified that *for each distribution, the distance of the synthetic day (from the original) is within this confidence interval.*

Finally, we applied the profiling methodology described in Section 4 on the synthetic subscribers. By doing so, we classify them and compare the per-class traffic behavior with the one created from the original dataset. Fig. 10(c) depicts the per-class behavior for the volume of traffic per session for the classified synthetic subscribers. It is possible to see that *this result is coherent with the one obtained for the original dataset* presented in Fig. 7(b) in terms of hourly behavior. For instance, the behavior for peak and non-peak hours is well defined and similar to the one from the original trace.

6. Discussion

In this section, we discuss some issues we judge interesting in the presented work. An important aspect in temporal network usage or protocol analysis is to be able to evaluate designed network solutions under the load it will be subjected to in realistic deployment. Subscribers with different profiles impose, on certain cases, totally different demands to the network. As extreme examples, our dataset shows that the heaviest user generates 22 million times more traffic than the lightest one. Moreover, the 276 thousand lightest subscribers generate a similar amount of traffic as the heaviest subscriber in an entire day. Still, spatial analysis of traffic demand is an important factor. Unfortunately, we did not have access to location information in the used dataset in order to perform such analysis. On the other side, it is worth to stress that our analysis (1) is performed on an individual-basis (instead of grouped- or network-basis) and (2) relies on previous work attesting the cyclical routine of mobile subscribers [2, 37]. These two particularities bring the flexibility and give us the possibility of combining, according to a temporal dimension, our generated synthetic traffic to a mobility trace describing spatial movements of subscribers over time. This would generate a new trace with subscribers generating traffic at the routinary locations where they usually are present at that time. Although this combination would allow the performance evaluation of network solutions under realistic spatio-temporal traffic demand, we argue a deeper modeling on spatial-temporal correlation of traffic demand is still strongly required (which is out of scope of this paper) and constitutes an open research in the literature (most probably due to the lack of available detailed traces combining traffic demand and mobility information).

Traffic demand is generally described by the set of different traffic parameters that characterize the demands of the users to the network. In this work, we have explored a set of parameters such as inter-arrival time, session duration, number of sessions, and volume of traffic. Each of these parameters were deeply assessed in our previous sections, but it is also interesting to examine the relations between them.

Fig. 10(d) shows a heatmap (better seen in colors) of the Pearson's correlation between those traffic parameters for all subscribers in all profiles. The intensity of the color on each cell of the matrix indicates the strength of negative or positive correlations. It is possible to see that the correlation between number of sessions and inter-arrival time goes from a low positive value on

LO to a high negative value on HF. Indeed, the correlation between them is 22%, -14% -26% -37% -45% -55% for LO, MO, HO, LF, MF, and HF, respectively. *It means that classes in which subscribers generate more sessions have higher negative correlation with the inter-arrival time. In general, the more sessions a user generates, the shorter they need to be to fit in a certain period of time.* A caveat here is that a user that generates few sessions could generate them in bursts, or sparsely separated in time. The former would result in small IAT and the latter in a larger IAT. For example, a large IAT of one hour is likely to be done for a user with few sessions per day, than a user with, for example, 300 sessions. In the same way, a small IAT could be generated for a user with both high or low number of sessions.

Another important aspect is the relation between volume of traffic and session duration. LO, MO, and HO classes present 13%, 14%, and 26%, respectively, i.e., a growing positive correlation with the session duration. LO, MO, and HO have, on average, 663, 6554 and 18624 seconds of session duration and 5090, 165214 and 6117322 KB of average session volume, respectively. *The growth of those metrics from one Occasional class to the next is due to the necessary increase on the session duration in order to accommodate the volume of traffic, considering that there is no significant raise on the number of sessions from LO, to MO, or HO.*

Finally, it is important to mention the correlation between number of sessions and volume of traffic. The correlation is overall low and positive between these two metrics for all profiles, but its behavior differs completely from Occasional to Frequent users. LO, MO, and HO have 29%, 17% and 0.4% correlation between number of sessions and volume of traffic, respectively, i.e., a decrease from LO, to MO and to HO. It happens because LO users have few sessions and low traffic volume, while MO and HO classes have significantly higher volume of traffic, but still few sessions. Therefore, the correlation is lower for MO and HO than for LO. Differently, LF, MF, and HF have 9%, 11% and 12% correlation between number of sessions and volume of traffic, respectively, i.e., a growth from LF, to MF and to HF. That is due to HF presenting both high volume and high number of sessions, while LF and MF present lower volume of traffic, but still high number of sessions.

Understanding network demands from users traffic parameters and their correlations is one of the contributions of our work. Moreover, *this work provides distributions to model workload characteristics of mobile subscribers' traffic demands and a framework to create a traffic generator based on these characterizations.* Our work proposes a general methodology, which can be used to extend the traffic generator to other traffic metrics, such as transfer rate and session duration. Therefore, it has implications in areas related to the design of new applications and network protocols (e.g., routing protocols), to the analysis on temporal network resource usage, or to network planning such as hotspot deployment [2]. as well as to the analysis on temporal network resource usage.

In this latter area, for instance, the objective is to provide the best placement for hotspots respecting certain constraints. For instance, one may desire to deploy a fixed amount of hotspots to maximize the amount of data offloaded from the network. The literature frequently presents the evaluation of hotspot deployment based on mobility datasets describing subscribers' trajectories. Although literature provides some mobility datasets, to the best of our knowledge none of them provides information of both mobility and traffic demands. Our traffic generator could be attached to the mobility datasets and it would allow to better exploit them. Considering the existence of a model that describes the mobility of a user by a set of timestamped locations he visits, we can associate traffic and location by the timestamp. In addition, we believe that the human routinary behavior will also help on this location-traffic association. Our routine is time-oriented and that implies our location at each time instant of a day. Following this intuition, we argue that if time

induces location, traffic per time will also induces traffic per location. It requires a more detailed analysis to verify this intuition, which is not in the scope of this paper, however, we think that our approach can be considered as a first step toward this direction.

We would like to point out mobile data traffic is heavily affected by the available pricing mode (both the pricing structure and price level) [Eduardo, can you please find 1 reference here?] and clearly evolves according to technology (e.g., 3G to 4G). In this context, unfortunately, we did not have any information on subscribers' data plans, and consequently, could not investigate the impacts brought by the pricing mode. This is however, a very interesting direction to follow. In addition, our dataset and resulting traffic generator describes subscribers' data traffic in a 3G cellular network only.

Another important aspect of the synthetic traffic generator is that it preserves the privacy of the original subscribers from whom the measurements came from. The non-existence of personal data attached to synthetic users allows us to limitlessly share our observations with the community without the necessity of sharing sensitive information inherent to datasets. One may argue that it is possible to anonymize the users identity, but literature shows that many attempts on that direction fail on protecting users privacy [38]. As shown in our analysis, our synthetic users generate traffic consistent with the original dataset and, thus strongly minimizes privacy issues. Considering extreme privacy concerns, e.g., one may argue that since the synthetic trace is created from real probability distributions, there is an information leakage. *Due to the fact that our model does not deal with a particular user, but with profiles that represent several users, that leakage is minimal compared to a scenario in which the original dataset is used.*

7. Conclusions and Next Steps

In this paper, we have first presented a characterization of a 4-month dataset that contains more than 1.05 billion data sessions from about 6.8 million smartphone users. Moreover, we propose a framework that automatically classifies those users according to their traffic demands into a limited number of profiles. Our approach takes advantage of repetitive users behavior due to their daily routines. Furthermore, we provide distributions that describe their traffic demands into peak and non-peak hours. Finally, from these distributions we create a traffic generator and evaluate the synthetic trace it generates. Our results show that the synthetic trace presents a consistent behavior when compared to original dataset.

As future work, we aim to model sessions' transfer rate and duration. Moreover, we intent to study the existence of real-world aspects on the synthetic trace other than the inter-arrival time, e.g., temporal auto-correlations of each measure. Additionally, we intend to apply and evaluate our traffic generator on different problems such as network planning (e.g., hotspot deployment). Relying on a future availability of geographic data, we plan to study the traffic parameters' spatial correlation.

References

[1] Cisco, Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018 (Feb. 2013).

[2] E. M. R. Oliveira, A. C. Viana, From routine to network deployment for data offloading in metropolitan areas, in: Proc. of IEEE SECON, 2014.

[3] J. Candia, M. Gonzalez, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabasi, Uncovering individual and collective human dynamics from mobile phone records, Journal of Physics A: Mathematical and Theoretical 41.

[4] R. Becker, R. Caceres, K. Hanson, J. Loh, S. Urbanek, A. Varshavsky, C. Volinsky, A tale of one city: Using cellular network data for urban planning, IEEE Pervasive Computing 10 (4) (2011) 18–26.

[5] J. Wortham, Cellphones now used more for data than for calls, New York Times.
[6] D. Naboulsi, R. Stanica, M. Fiore, Classifying call profiles in large-scale mobile traffic datasets, in: Proc. of IEEE Infocom, 2014.
[7] A. Pawling, N. V. Chawla, G. Madey, Anomaly detection in a mobile communication network, Computational and Mathematical Organization Theory 13 (4) (2007) 407–422.
[8] S. Hoteit, S. Secci, Z. He, C. Ziemlicki, Z. Smoreda, C. Ratti, G. Pujolle, Content consumption cartography of the paris urban region using cellular probe data, in: Proc. of the 1st Workshop on Urban Networking (ACM UrbaNe), 2012.
[9] C. W. O. O. A. Abidogun, A self organizing maps model for outlier detection in call data from mobile telecommunication networks, in: Proc. of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC), 2004.
[10] R. A. Becker, R. Cceres, K. Hanson, J. M. Loh, S. Urbanek, E. Varshavsky, C. Volinsky, Clustering anonymized mobile call detail records to find usage groups, Workshop on Pervasive and Urban Applications (PURBA) (2011).
[11] A. Stoica, Z. Smoreda, C. Prieur, J.-L. Guillaume, Age, Gender and Communication Networks, in: V. Blondel, G. Krings (Eds.), NetMob 2010 Workshop on the Analysis of Mobile Phone Networks, 2010.
[12] A. Mehrotra, A. Nguyen, J. Blumenstock, V. Mohan, Differences in phone use between men and women: Quantitative evidence from rwanda, in: Proceedings of the Fifth International Conference on Information and Communication Technologies and Development, ICTD '12, ACM, 2012, pp. 297–306.
[13] K. Lee, J. Lee, Y. Yi, I. Rhee, S. Chong, Mobile data offloading: How much can wifi deliver?, Networking, IEEE/ACM Transactions on 21 (2) (2013) 536–550.
[14] R. Keralapura, A. Nucci, Z.-L. Zhang, L. Gao, Profiling users in a 3g network using hourglass co-clustering, in: Proc. of ACM MobiCom, 2010.
[15] A. Vaccari, L. Liu, A. Biderman, C. Ratti, F. Pereira, J. Oliveirinha, A. Gerber, A holistic framework for the study of urban traces and the profiling of urban processes and dynamics, in: Proc. of Int. IEEE Conf. on Intelligent Transportation Systems (ITSC), 2009.
[16] P. Paraskevopoulos, T. C. Dinh, Z. Dashdorj, T. Palpanas, L. Serafini, Identification and characterization of human behavior patterns from mobile phone data, in: Proc. of NetMob, 2013.
[17] R. M. Pulselli, P. Romano, C. Ratti, E. Tiezzi, Computing urban mobile landscapes through monitoring population density based on cellphone chatting, Int. Journal of Design and Nature and Ecodynamics 3.
[18] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, C. Ratti, Towards estimating the presence of visitors from the aggragate mobile phone network activity they generate, in: Proc. of Intl. Conference on Computers in Urban Planning and Urban Management, 2009.
[19] Q. Lin, Mobile customer clustering analysis based on call detail records, in: Communications of the IIMA, Vol. 7, 2007.
[20] Alcatel-Lucent, Alcatel-lucent 9900 wireless network guardian, White Paper (Dec. 2012).
[21] M. Chuah, W. Luo, X. Zhang, Impacts of inactivity timer values on umts system capacity, in: Proc of. IEEE WCNC, Vol. 2, 2002, pp. 897–903 vol.2.
[22] U. Paul, A. Subramanian, M. Buddhikot, S. Das, Understanding traffic dynamics in cellular data networks, in: Proc. of IEEE Infocom, 2011.
[23] D. B. Carr, A. R. Olsen, D. White, Hexagon mosaic maps for displaying univariate and bivariate geographical data, Cartography & Geographical Information Systems 19 (1992) 228–236.
[24] M. Shafiq, L. Ji, A. Liu, J. Pang, J. Wang, Large-scale measurement and characterization of cellular machine-to-machine traffic, Networking, IEEE/ACM Transactions on 21 (6) (2013) 1960–1973.
[25] J. Brea, J. Burroni, M. Minnoni, C. Sarraute, Harnessing mobile phone social network topology to infer users demographic attributes, in: Proceedings of the 8th Workshop on Social Network Mining and Analysis, SNAKDD'14, ACM, 2014.
[26] International Trade Union Confederation, Frozen in time: Gender pay gap unchanged for 10 years, Tech. rep. (2012).
[27] M. Halkidi, Y. Batistakis, M. Vazirgiannis, On clustering validation techniques, Journal of Intelligent Information Systems 17 (2-3) (2001) 107–145.
[28] R. R. Sokal, C. D. Michener, A statistical method for evaluating systematic relationships, University of Kansas Scientific Bulletin 28 (1958) 1409–1438.
[29] G. Milligan, M. Cooper, An examination of procedures for determining the number of clusters in a data set, Psychometrika 50 (2) (1985) 159–179.
[30] E. M. R. Oliveira, A. C. Viana, K. P. Naveen, C. Sarraute, Measurement-driven mobile data traffic modelling in a large metropolitan area, Tech. rep., INRIA (2014).
URL https://hal.inria.fr/hal-01073129v4/document
[31] T. Hossmann, T. Spyropoulos, F. Legendre, Know thy neighbor: Towards optimal mapping of contacts to social graphs for DTN routing, in: Proc. of IEEE INFOCOM, 2010.
[32] R. B. D'Agostino, M. A. Stephens, Goodness-of-Fit-Techniques, Vol. 68, CRC Press, 1986.

[33] K. Pearson, X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philosophical Magazine Series 5 50 (302) (1900) 157–175.

[34] A. Bhattacharyya, On a measure of divergence between two statistical populations defined by their probability distributions, Bulletin of the Calcutta Mathematical Society 35 (1943) 99–109.

[35] D. Comaniciu, V. Ramesh, P. Meer, Kernel-based object tracking, IEEE Trans. on Pattern Analysis and Machine Intelligence 25 (5) (2003) 564–577.

[36] S. Kullback, R. A. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 49–86.

[37] E. Mucceli, A. C. Viana, C. Sarraute, J. Brea, I. Alvarez-Hamelin, On the regularity of human mobility, To appear in Pervasive and Mobile computing (PMC) Journal, Elsevier.

[38] President's Council of Advisors on Science and Technology, Big Data and Privacy: A Technological Perspective, Tech. rep., Executive Office of the President (5 2014).